Journal of Biomedical Informatics 42 (2009) 726–735

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Literature mining on pharmacokinetics numerical data: A feasibility study

Zhiping Wang [a], Seongho Kim [a], Sara K. Quinney [a], Yingying Guo [b], Stephen D. Hall [b], Luis M. Rocha [c,d], Lang Li [a,*]

[a] Division of Biostatistics, Department of Medicine, School of Medicine, Indiana University, 410 West 10th Street, Suite 3044, Indianapolis, IN 46202, USA
[b] Eli Lilly and Company, Indianapolis, IN, USA
[c] School of Informatics, Indiana University, Bloomington, IN, USA
[d] Instituto Gulbenkian de Ciencia, Oeiras, Portugal

ARTICLE INFO

Article history:
Received 1 December 2008
Available online 2 April 2009

Keywords:
Clearance
Data mining
Entity recognition
Information extraction
Linear mixed model
Midazolam
Pharmacokinetics

ABSTRACT

A feasibility study of literature mining is conducted on drug PK parameter numerical data with a sequential mining strategy. Firstly, an entity template library is built to retrieve pharmacokinetics relevant articles. Then a set of tagging and extraction rules are applied to retrieve PK data from the article abstracts. To estimate the PK parameter population-average mean and between-study variance, a linear mixed meta-analysis model and an E–M algorithm are developed to describe the probability distributions of PK parameters. Finally, a cross-validation procedure is developed to ascertain false-positive mining results. Using this approach to mine midazolam (MDZ) PK data, an 88% precision rate and 92% recall rate are achieved, with an *F*-score = 90%. It greatly out-performs a conventional data mining approach (support vector machine), which has an *F*-score of 68.1%. Further investigate on 7 more drugs reveals comparable performances of our sequential mining approach.

1. Introduction

In recent decades, a new drug requires an average of 15 years and approaching a billion dollars in research and development. Unfortunately, only one in 10 drugs that enter clinical testing receives eventual FDA approval [1]. Scientists have become increasingly mechanistic in their approach to drug development [2]. The recent ability to integrate genetic mutations and altered protein expression to pharmacokinetics (PK) and pharmacodynamic (PD) models allow a deeper understanding of the mechanisms of disease and therapies that are genuinely targeted [3–6]. In 2004, the FDA released a report entitled: "Innovation or Stagnation, Challenge and Opportunity on the Critical Path to New Medical Products" [7]. Among its six general topic areas, three of them emphasized the importance of computational modeling and bioinformatics in biomarker development and streamlining clinical trials [8,9]. In multiple follow-up papers, clinical researchers, experimental biologists, computational biologists, and biostatisticians from both academia and industry all cheered the FDA leadership in this critical path, and pointed out the challenges and opportunities of the PK/PD model based approach in drug development [10–13].

To fulfill the PK/PD modeling potential in drug development, there is an enormous need for pharmacology database of PK/PD parameters. For example, to specify the first human dose of a new compound, based on animal studies, one needs available *in-vitro* and *in-vivo* PK parameters from its comparators (in market) [10,13]. However, these PK data are rarely available from public pharmacology databases. DiDB (http://www.druginteractioninfo.org/) has manually accumulated published PK publications for each drug. However, DiDB has no summarized PK parameters. DrugBANK (http://redpoll.pharmacy.ualberta.ca/drugbank/) is a comprehensive pharmacology database. It has collected rich annotations on drug's chemistry, structure, mechanism, pathway, and targets, but has very sparse PK data. The other one is PharmGKB (http://www.pharmgkb.org/). It stores enormous amount of pharmacogenetics (PG) data from ongoing PG studies sponsored by National Institute of General Medical Science (NIGMS), but very limited PK data are available.

The shortage of manually maintained drug databases is the lack of ability to ensure the completeness and timely update of PK parameters. The widely used online literature search service, PubMed, contains about 18 million abstracts from MEDLINE and additional life sciences journals. Its collection grows at 40,000 new biomedical abstracts every month [14]. Even for some subsections of the drug literature in PubMed, e.g. the drug therapy review articles which grow about 10,000 English-language articles per year [15]. Therefore, it is not possible to keep track of all relevant literature manually. In addition, the complexity of information from PK/PD studies makes it even harder for the data processing when various drug doses, administration routes, patients, sample collection intervals, and the like are involved into the collected

* Corresponding author. Fax: +1 317 274 2678.
E-mail address: lali@iupui.edu (L. Li).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.03.010

information. Moreover, the drug PK information that needs to be collected depends heavily on the mechanistic of PK/PD models and their simulations, which is in turn driven by the science. Thus the required literature information becomes a moving target. Consequently, the manual knowledge base accumulation approach cannot meet all the challenges.

One effective alternative is literature mining [16,17], which trains the machine to discover useful information or make novel hypothesis on publications. Its machine learning methodology has the advantage of processing large amount of information within a short time, the flexibility of adaptation and integration with follow-up applications, e.g. PK/PD model developments. This technique has been applied on many biomedical research problems [16], and has been proved sufficiently accurate to be used in practice [17]. However, there is no single strategy working equally well for all types of information extraction requirements, and little research has been done for the numerical pharmacokinetic data extraction from scientific publications. This situation calls for the development of a novel strategy specifically to extract the numerical PK/PD parameters.

Literature mining on PK parameters is highly unique. Firstly, important PK parameters (entities) are specifically defined, e.g. absorption rate, bioavailability, clearance, etc. These PK parameters are usually available from different drug studies, which may vary by factors such as units, sub-populations, study designs, and dose regimens. Thus a set of experience based standards need to be created to normalize the mined data. Secondly, the retrieved information can be incomplete and unbalanced from different published paper. So the missing information needs to be estimated and imputed from the known PK parameters according to their relationships. Thirdly, false positive findings need to be cleaned from mined results as thorough as possible.

One barrier in the study of literature mining is the relative lack of standards to evaluate the performance of mining strategies. This situation stimulates the generation of some ongoing evaluation resources and benchmarks, e.g. Knowledge Discovery and Data Mining (KDD) Challenge Cup [19], TREC Genomics Track [20], BioCreAtIvE [21], etc. However, these evaluation resources are usually designed for other biomedical problems. As for the application of numerical data targeted literature mining in the field of PK/PD study, little previous work is available for reference. Therefore, one goal of our research is to establish a new validation data set to offer to the community.

Additionally, in this paper, we test a novel literature mining strategy we have developed. Our approach targets drug PK parameter numerical data extraction. It possesses entity recognition, information extraction, and outlier detection. In particular, we employ a likelihood based statistical model to describe the distribution of PK parameters, develop an expectation-maximization (EM) data evaluation algorithm to estimate the PK parameter population means and variances, and an outlier detection rule to remove false-positive mining results. The details of the mining and evaluation methodology are presented in Section 2. The data mining implementation for drug, midazolam (MDZ), is illustrated in Section 3. Conclusions are reached in Section 4.

2. Methods

2.1. MDZ case-study overview

The goal of our literature mining approach is to extract all pharmacokinetics (PK) related information for a given drug. In this paper, we use midazolam (MDZ) as the test drug. The mining is based on abstracts from PubMed. One example of a MDZ PK relevant abstract is provided below [22].

> To study the effects of cirrhosis of the liver on the pharmacokinetics of midazolam single IV (7.5 mg as base) and p.o. (15.0 mg as base) doses of midazolam were administered to seven patients with cirrhosis of the liver and to seven healthy control subjects. The elimination of midazolam was significantly retarded in the patients as indicated by its lower total clearance (3.34 vs. 5.63 ml/min/kg), lower total elimination rate constant (0.400 vs. 0.721 h^{-1}), and longer elimination half-life (7.36 vs. 3.80 h). The bioavailability of oral midazolam was significantly ($P < 0.05$) higher in patients than controls (76% vs. 38%).

The search engine of PubMed is not subtle enough to limit the search results to a specific topic, i.e. human PK study. So a further filtering step is necessary to remove irrelevant articles from PubMed search results, and keep the PK relevant abstracts which usually contain information from the following relevant keyterm categories.

- Subject type (race, age, sex, etc.) and size.
- MDZ dose and administration route (oral, intravenous, etc.).
- PK parameters, such as AUC (area under the concentration–time curve), half-life, bioavailability, clearance, etc.

Besides the keyterm categories above, we limit the mining to PK data from healthy human subjects and the target drug only (i.e. no other factors involved such as drug inhibitor/activator) to comply with the requirements of drug PK study.

Hence, in the example abstract, the literature mining tool should be able to extract "seven healthy control subjects" as subject, "IV (7.5 mg as base) and p.o. (15.0 mg as base)" as dose, "3.34 vs. 5.63 ml/min/kg" as total clearance, "0.400 vs. 0.721 h^{-1}" as elimination rate, "7.36 vs. 3.80 h" as half-life, and "76% vs. 38%" as bioavailability. To be more precise, the mining tool should be able to recognize which value is for drug MDZ in the comparison situation, e.g. for the total clearance data, "3.34 ml/min/kg" is from patients and "5.63 ml/min/kg" is from healthy subjects.

To achieve the goal above, we developed a rule-based information extraction system. The architecture is shown in Fig. 1. The abstracts are downloaded from PubMed after an initial query for a target drug, e.g. midazolam. Abstract texts are preprocessed such that they are divided into sentences, and different forms of the same terms are trimmed. The next step is entity recognition. It determines sentence relevance, and tags the trimmed sentence terms as various entity classes. At the end of this step, only the more relevant abstracts are left and well tagged. In the information extraction step, a set of extraction rules are manually created and implemented. Then the mined data are analyzed by a statistical model to detect and remove outliers, which are potential false positive data. The final data set is saved into a PK parameter database.

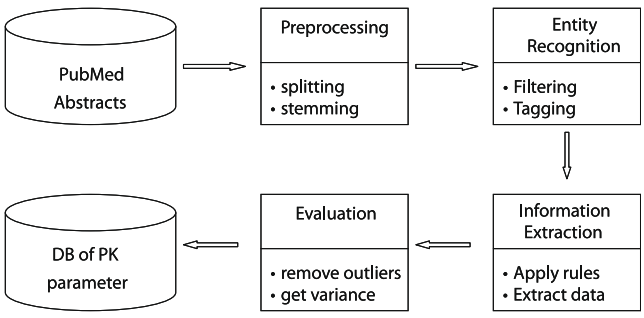

Fig. 1. The architecture of literature mining tool.

2.2. Text preprocessing

Our PubMed search uses the drug name, e.g. midazolam, as the unique keyterm in a query. The search results are downloaded with the XML format to get the structured abstract information. In the following mining process, only article title (⟨ArticleTitle⟩), abstract (⟨AbstractText⟩) and paper type (⟨PublicationType⟩) information is utilized from the XML format abstract.

The goal of the preprocessing step is to split the abstract text into units of sentences. There are some existing tools to do this job (e.g. SentenceDetector [23], MxTerminator [24]). Considering the simple grammar of the abstracts, we applied a Perl module (Lingua::EN::Sentence) for sentence splitting. The Porter stemming algorithm [25] is used to deal with the common morphological and inflexional endings from words in English. After stemming, each word in the abstracts is normalized into a standard form.

2.3. Entity recognition

2.3.1. Entity template library

Keyterm based PubMed search usually returns a large number of abstracts, e.g. 7129 midazolam abstracts. To increase the precision of the mined results, an abstract filtering step is necessary after the text preprocessing. Firstly, as we limit the mining of PK data from healthy human subjects, the studies on diseased subjects should be removed. The human subject information (health status, race, weight, etc.) is highly important and usually reported in pharmacokinetics studies. Most article abstracts state clearly whether the human subjects are healthy or diseased (patients). So if one abstract only mentions patients or diseased subjects, it is usually irrelevant; but if there is co-existence of healthy subject information (this is usually the control in clinical studies), it is still considered as subject relevant. For abstracts without any subject information, we kept them as relevant in case of data loss. Secondly, an entity template library is built based on expert knowledge for the further abstract filtering. It summarizes key factors in determining an abstract's relevance. Table 1 is a library example, which contains a list of relevant keyterms and a list of forbidden terms. The terms are in the stemmed format. Because some relevant abstracts do not have human subject information, subject terms are not included in the keyterm list. Thus, these articles can be kept as relevant for future full text mining purposes. In addition, the drug related terms should correspond to the studied drug. For midazolam, such terms are "midazolam", "mdz", "cyp3a", and "p450".

Table 1
Keyterms and forbidden terms.

Keyterms		Forbidden terms	
ROUTE	orally	NTITLE	mice
	intraven		mouse
	administr		rat
	i.v.		animal
	intramuscular		penguins
DRUG	⟨Drug terms⟩		pig
			horse
			human liver
			microsom
			review
PK	clearance	NTYPE	review
	pharmacokinet		
	concentr		
	bioavail		
	auc		
	elimin		
	c(max)		
	half-lif		

The entity template library is a representation model for the relevant abstracts. The PubMed search abstracts are further filtered by this library. An abstract is considered relevant if it contains at least one term from each of the keyterm categories, which include drug administration routes (ROUTE), PK parameters and DRUG (Table 1); and the abstract is considered as irrelevant if it contains one or more forbidden terms in either ⟨NTITLE⟩ or ⟨NTYPE⟩ (Table 1). As MDZ is primarily a CYP3A substrate, all of its DRUG keyterms are related to this metabolic enzyme. The ⟨NTITLE⟩ is the forbidden term list for article titles. These terms mostly represent animal and *in-vitro* studies (Table 1). The other forbidden term, ⟨NTYPE⟩, is used to recognize the review articles. Since review articles contain PK data only from other publications, they do not provide additional information.

2.3.2. Tagging entities

All information in the keyterm categories is necessary for a drug PK study. In order to extract all the PK data from the abstract, it is critical to properly tag all these relevant terms in each sentence.

2.3.3. Subject tagging

The subject information usually contains all or part of the following four key components: size, description, race and subject types (Table 2). ⟨SUB_part: term⟩ is used to represent a term in each component, e.g. "seven healthy control subjects" can be tagged as "⟨SUB_N: seven⟩ ⟨SUB_D: healthy⟩ control ⟨SUB_T: subjects⟩". The subject tagging starts from searching the TYPE component in one sentence, then trace back to the other components. If at least two components exist in one sentence, they are tagged as subject information.

2.3.4. Drug tagging

Most PK study related abstracts cover multiple drug names, e.g. drug–drug interaction studies. PK parameters in these abstracts are available for multiple drugs. The mining tool should recognize different drug names. A complete list of all the approved drugs from the U.S. Food and Drug Administration (FDA) website is downloaded, and a drug name dictionary is built from this list. Thus, the drug entities in the abstracts can be correctly tagged, e.g. midazolam to ⟨DRUG: midazolam⟩.

2.3.5. Dosing tagging

The dosing tagging information is illustrated in Table 3. The first column shows different drug administration ways, and the second column lists all possible units for dosing. The dose is located by searching the numerical data lying ahead of its unit. In sentences, the administration routes and units after numerical data are important dosing tags. As these tags are highly compact, they usually occur together. For example, the following two dosing related text segments

- Midazolam oral (15 mg) and intravenous (0.05 mg kg^{-1}) was given

Table 2
Subject related terms.

Number (N)	Description (D)	Race (R)	Type (T)
[0–9]	Young	Chinese	Volunteers
{one...ten}	Healthy	Japanese	Subjects
{eleven... twenty}	Medication-free	Vietnamese	Individuals
	Elderly	European	Women
{thirty... ninety}	Male	Caucasian	Men
	Female	Mexican	Americans
	White	American	Immigrants
	Premenopausal	African	
	Nonsmoking		

Table 3
Dosing information.

Administration (A)	Unit (U)
Oral/orally	mg
iv/i.v.	−mg
po/p.o.	μg
s.c.	mg kg(−1)
Infuse	mg(−1) kg
Intramuscular	mg/kg
Intravenous/intravenously	mg kg^{-1}

- 7.5 mg dose of midazolam was given orally
  are tagged as
- ⟨DRUG: Midazolam⟩ ⟨Dose_A: oral⟩ (15 ⟨Dose_U: mg⟩) and ⟨Dose_A: intravenous⟩ (0.05 ⟨Dose_U: mg kg^{-1}⟩) was given
- 7.5 ⟨Dose_U: mg⟩ dose of ⟨DRUG: Midazolam⟩ was given ⟨Dose_A: orally⟩.

2.3.6. PK parameter tagging

Drug clearance is chosen as the test PK parameter data mining performance, since it has comparably more numerical data available in the abstracts. The important tags for the clearance relevant value and unit are,

- Clearance terms (T): [systemic/oral] clearance.
- Value (V).
- Unit (U) examples: ml/min/kg; l/kg/h; ml/min; l/h.

As there are two types of clearance, systemic clearance and oral clearance, a type classification is needed in the following data analysis step. The clearance value is usually reported in both sample mean and standard deviation. The co-existence of the clearance keyterms and units is a unique identification, and the tagging is done by identifying them together in one sentence. For example, the phrase "the systemic clearance of midazolam was unchanged (37.7 ± 11.3 l/h)" is tagged as "the ⟨CLR_T: systemic clearance⟩ of ⟨DRUG: midazolam⟩ was unchanged (⟨CLR_V: 37.7 ± 11.3⟩ ⟨CLR_U: l/h⟩)".

After the tagging process, the relevant elements in each sentence are recognized. The tagged sentences in each abstract are kept for the following information extraction. All the untagged sentences are removed.

2.4. Information extraction

Information extraction is to extract the information from three prescribed tagging items: dosing, subject, and PK parameters. The subject and dosing information can be extracted easily given a well tagged sentence. For example, given the tagged phrase "⟨SUB_N: seven⟩ ⟨SUB_D: healthy⟩ control ⟨SUB_T: subjects⟩", the machine easily locates the subject information. Similarly, the tagged phrase "7.5 ⟨Dose_U: mg⟩ dose of ⟨DRUG: Midazolam⟩ was given ⟨Dose_A: orally⟩" clearly shows "7.5 mg orally" as the dosing information for midazolam. The tagged sentence "⟨DRUG: Midazolam⟩ ⟨Dose_A: oral⟩ 15 ⟨Dose_U: mg⟩ and ⟨Dose_A: intravenous⟩ 0.05 ⟨Dose_U: mg kg^{-1}⟩ " indicates a simple sequential parsing of information for oral dosing and intravenous dosing as "oral 15 mg; intravenous 0.05 mg kg^{-1}".

PK parameter data extraction is more complicated. As multiple drugs are usually involved into the PK studies, one abstract sentence may contain PK data for both target drug and other drugs. Even if one sentence discusses the target drug only, the data can reflect its PK value change caused by other study drugs. The following sentence reflects this complexity.

> Rifampin significantly ($P < 0.0001$) increased the systemic and oral clearance of midazolam from 0.44 ± 0.2 L h/kg and 1.56 ± 0.8 L h/kg to 0.96 ± 0.3 L h/kg and 34.4 ± 21.2 L h/kg, respectively.

Two drugs, midazolam and rifampin, are mentioned in this sentence, and the clearance values contain both control and affected cases. The information extraction needs to make the correct decision that this sentence discusses midazolam, but not rifampin; and the control clearance values come first (0.44 ± 0.2 for systemic clearance; 1.56 ± 0.8 for oral clearance). There are two steps to discriminate the target drug. First, if the title or the occurrence frequency of term "midazolam" shows strong signal that the abstract is about midazolam but not rifampin, this sentence is most likely to be midazolam. Secondly, it is still possible that one clearance value is for rifampin for the sake of comparison. In order to deal with this case, a set of extraction rules are created. The rules are explained in detail in the follow up example. After the tagging step, this sentence example is converted to

> ⟨DRUG: Rifampin⟩ significantly ($P < 0.0001$) ⟨CHG: increased⟩ the ⟨CLR_T: systemic⟩ and ⟨CLR_T: oral clearance⟩ of ⟨DRUG: midazolam⟩ from ⟨CLR_V: 0.44 ± 0.2⟩ ⟨CLR_U: L. h/kg⟩ and ⟨CLR_V: 1.56 ± 0.8⟩ ⟨CLR_U: L h/kg⟩ to ⟨CLR_V: 0.96 ± 0.3⟩ ⟨CLR_U: L h/kg⟩ and ⟨CLR_V: 34.4 ± 21.2⟩ ⟨CLR_U: L h/kg⟩, respectively.

The tag "⟨CHG⟩" is an important one to show the change of clearance value caused by the co-existence of other drugs. Hence, the "increased" case of ⟨CHG⟩ tag, the smaller value of clearance data is usually the control, i.e. study with no drug interaction effect, which should be extracted. Now the simple representation pattern for this sample sentence is "⟨DRUG1⟩ ⟨CHG⟩ ⟨CLR_T1⟩ ⟨CLR_T2⟩ ⟨DRUG⟩ ⟨CLR_V1⟩ ⟨CLR_V2⟩ ⟨CLR_V3⟩ ⟨CLR_V4⟩". The rules to extract clearance information for this type of pattern are listed below:

- Find clearance type ⟨CLR_T1⟩ ⟨CLR_T2⟩.
- Find value change type ⟨CHG⟩.
- Each value change involves two clearance values for one clearance type, hence there should be four clearance values (⟨CLR_T1_V1⟩ to ⟨CLR_V4⟩).
- The clearance values for ⟨CLR_T1⟩ can be (⟨CLR_V1⟩ ⟨CLR_V2⟩) or (⟨CLR_V1⟩ ⟨CLR_V3⟩). Choose the pair with the smaller difference, and the smaller value in that pair is ⟨CLR_T1⟩. For example, the systemic clearance is 0.44 ± 0.2 L h/kg.
- The other two values are for ⟨CLR_T2⟩. Similarly, the smaller value of the two is chosen for ⟨CLR_T2⟩. For example, the oral clearance is 1.56 ± 0.8 L h/kg.

These extraction rules cover regular expressions of clearance data, considering single and multiple drug occurrences, different clearance types, and clearance value changes.

2.5. Evaluation – linear mixed model meta-analysis for PK parameter estimation and outlier detections

Because the mined PK parameter numerical data may contain some false positive values, an evaluation mechanism is needed to remove them as outliers. The population mean and variance of PK parameters are also requested to be estimated. We developed a linear mixed model meta-analysis approach for this purpose. The PK parameter values are assumed to follow the normal distribution as illustrated in Eq. (1).

$$\bar{\theta}_{.k} \sim N(\theta_k, se_k^2) \quad \theta_k \sim N(\theta, \sigma^2) \tag{1}$$

The first normal distribution is at the study level, in which $\bar{\theta}_{\cdot k}$ (the sample mean of study k) has study-specific mean θ_k, sample standard error se_k^2, where $k = 1, \ldots, K$, indicates the studies. The second normal distribution is at the population level, in which θ_k has the population mean θ, and σ^2 is its between-study variance. The population and study level PK parameters are two common statistics concepts in the pharmacokinetics meta-analysis literature [28]. The population PK parameter refers to its population-average mean, and a study-specific PK parameter refers to its sub-population mean, in which the study was sampled from. In this paper, we assume that PK data from one paper is a study, which is denoted by k.

In Eq. (1), $\bar{\theta}_{\cdot k}$ and se_k^2 are observed data from the literature mining results. The unknown parameters θ, σ^2 and θ_k are estimated by the following expectation and maximization algorithm. The expectation step estimates θ_k by Eq. (2).

$$\hat{\theta}_k = \left[\frac{1}{se_k^2} + \frac{1}{\sigma^2}\right]^{-1} \cdot \left[\frac{\bar{\theta}_{\cdot k}}{se_k^2} + \frac{\theta}{\sigma^2}\right] \tag{2}$$

The values of population mean θ and population variance σ^2 are estimated in the maximization step by Eq. (3). The E–M iterative procedure stops when the estimated values are stable.

$$\begin{aligned} &L(\theta, \sigma^2, \theta_k | \bar{\theta}_{\cdot k}, se_k^2) \propto \prod_k N(\theta_k, se_k^2) \cdot N(\theta, \sigma^2) \\ &\frac{\partial}{\partial \theta} \log L = 0 \Rightarrow \hat{\theta} = \frac{\sum_{k=1}^{n} \theta_k}{n} \\ &\frac{\partial}{\partial \sigma^2} \log L = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_{k=1}^{n} (\theta_k - \theta)^2}{n} \end{aligned} \tag{3}$$

Based on the meta-analysis, the standard error of the estimated population mean is expressed in Eq. (4),

$$se(\theta) = \sqrt{\frac{1}{\sum_{k=1}^{K} \frac{1}{(\sigma^2 + se_k^2)}}}. \tag{4}$$

2.6. Validation and classification

Some PK parameters have multiple types, but the abstracts do not always state clearly which type a numerical data refers to, e.g. some MDZ abstracts just use a single word "clearance" to represent either systemic clearance or oral clearance. In order to classify the unknown clearance type, the probability functions are established from known oral and intravenous clearance data with prescribed linear mixed model. Denote them as, $P[\cdot|\theta_{PO}, se_{PO}^2, \sigma_{PO}^2]$ and $P[\cdot|\theta_{sys}, se_{sys}^2, \sigma_{sys}^2]$, respectively. For an unknown type sample mean clearance value, $\bar{\theta}_{\cdot k,unknown}$ is classified by Eq. (5).

$$\begin{cases} P[\bar{\theta}_{\cdot k,unknown} | \theta_{PO}, se_{PO}^2, \sigma_{PO}^2] > P[\bar{\theta}_{\cdot k,unknown} | \theta_{sys}, se_{sys}^2, \sigma_{sys}^2] \Rightarrow type = PO, \\ P[\bar{\theta}_{\cdot k,unknown} | \theta_{PO}, se_{PO}^2, \sigma_{PO}^2] < P[\bar{\theta}_{\cdot k,unknown} | \theta_{sys}, se_{sys}^2, \sigma_{sys}^2] \Rightarrow type = Systemic. \end{cases} \tag{5}$$

Then a leave-one-out strategy is implemented to validate the classification results. For each classified data set, one single data is taken out and the rest go through the prescribed linear mixed model meta-analysis (Section 2.5). If this left-out data is 2.5 standard deviations from the population mean, it is considered as an outlier. To save the computation time, the data are ranked first, and this leave-one-out process is conducted iteratively from both the bottom and the top of the ranked data until the left-out data is not considered as outliers.

3. Results

3.1. MDZ data mining

3.1.1. Entity recognition

In this paper, midazolam (MDZ) is used to test our literature mining strategy. The keyterm "midazolam" in PubMed search returns over 7129 article records. After applying the entity template, out of the 7129 PubMed abstracts, 393 abstracts are considered as MDZ PK relevant. Among those 393 abstracts, 170 are truly relevant by the manual checking. The precision is 43%.

3.1.2. Information extraction

From 393 abstracts, the information extraction returns 53 abstracts. 43 out of 53 abstracts contain the true MDZ clearance data. Hence the precision improves to 81%. The same information extraction rules are also applied directly to the starting 7129 PubMed abstracts. It returns 120 abstracts, and a much lower precision, 36% (dashed lines in Fig. 2). This analysis shows the importance and the power of the entity template step.

3.1.3. Evaluation

Linear mixed model meta-analysis is implemented to classify the oral and systemic clearances, and remove the outlier data and abstracts. After this evaluation step, only 48 final abstracts are left, and 42 of them are true (precision 88%). The precision of

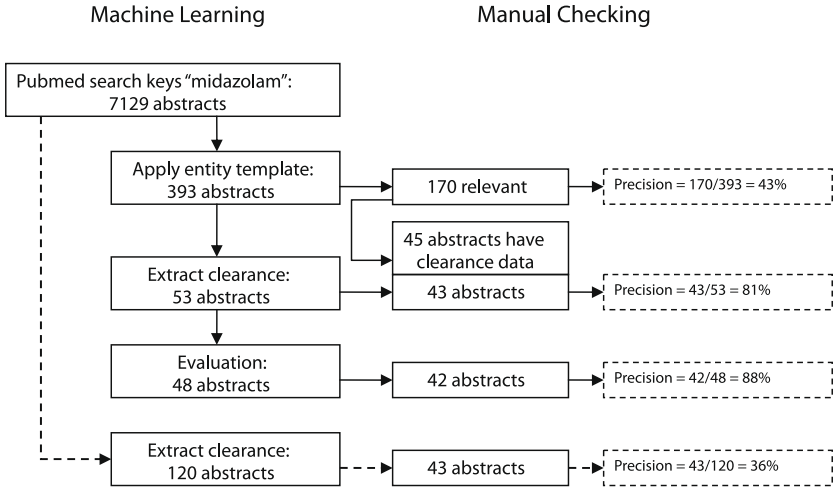

Fig. 2. Precision performance analysis of the machine learning algorithm in all MDZ related abstracts.

Table 4
Mined and Validated MDZ Clearance Data.

	Oral	Systemic	Unknown
Mined Clearance data (L/h)	*0.72*, *4.9*, *8.2*, **31.98**, 42.6, 43.2, *52.32*, 68.64, 84.78, 109.2, 116.8, 124.8, 137, 152, **215.9**, *1289*	15.12, 18.6, 22.98, 28, 32, 33.06, 33.6, 35.2, 36.9, 37.7, **77.28**, **84.78**	*0.81*, *1.14*, *2.016*, *2.11*, *2.26*, *2.4*, *3*, *4.6*, *5.58*, *6.6*, *14.94*, 15.9, *16.75*, 16.98, 19.02, 19.38, 19.5, *20.16*, 21.12, *21.5*, 22.2, *22.56*, 23.28, *23.3*, 23.4, *23.5*, *23.52*, 23.664, 23.94, 24, 24.8, 25.14, 25.2, *25.86*, 25.92, 27.024, 27.78, 28, 28.2, *28.8*, 28.96, 29.904, 30.12, 30.64, 32.16, 33.78, 34.08, 36.77, 36.96, 37.44, 37.92, 38.88, 39.17, 39.22, *40.8*, 42.4, 45.12, 45.6, 46.08, 51.2, 52.8, *53.8*, *54.6*, *54.72*, 58.56, 59.04, 59.2, 66.24, 78.6, 97.5, 99.36, *132*, *144*, 146, 166.56, *1281*, *2272*, *3328*, *5472*, *17616*
	Oral		Systemic
Clearance after evaluation (L/h)	42.4, 42.6, 43.2, 45.12, 45.6, 46.08, 51.2, 52.8, *53.8*, *54.6*, *54.72*, 58.56, 59.04, 59.2, 66.24, 68.64, 78.6, 97.5, 99.36, 109.2, 116.8, 124.8, *132*, 137, *144*, 146, 152, 166.56		*14.94*, 15.12, 15.9, *16.75*, 16.98, 18.6, 19.02, 19.38, 19.5, *20.16*, 21.12, *21.5*, 22.2, *22.56*, 22.98, 23.28, *23.3*, 23.4, *23.5*, *23.52*, 23.664, 23.94, 24, 24.8, 25.14, 25.2, *25.86*, 25.92, 27.024, 27.78, 28, 28, 28.2, *28.8*, 28.96, 29.904, 30.12, 30.64, 32, 32.16, 33.06, 33.6, 33.78, 34.08, 35.2, 36.77, 36.9, 36.96, 37.44, 37.7, 37.92, 38.88, 39.17, 39.22, *40.8*, 42.4, 45.12, 45.6, 46.08, 51.2

Note: The mined clearance data have three types: oral, systemic and unknown type. The false positive data was labeled *italics*; the false negative data which was removed in the validation step was labeled **bold**.

the mining goes from 43% in entity recognition to 81% in clearance data extraction, and reaches 88% after evaluation.

3.1.4. MDZ clearance parameter estimation and outlier detections

The MDZ PK clearance data from the information extraction are shown in the first row of Table 4. The mined clearance data have three types: oral clearance, systemic clearance and clearance with unknown mechanisms. The values are normalized based on an estimated average human body weight 80 kg, and verified by manually going through the abstracts. False positive clearance data are labeled in italics.

The mined clearance data then go through the linear mixed model meta-analysis to estimate the distributions for the systemic/oral clearance and remove the outliers. The calculated distributions are displayed in Fig. 3. The population mean ± se of systemic clearance is 27.8 ± 1.0 L/h, and its between-study standard deviation is 7.31; oral clearance is 78.1 ± 6.0 L/h, and its between-study standard deviation is 32.8.

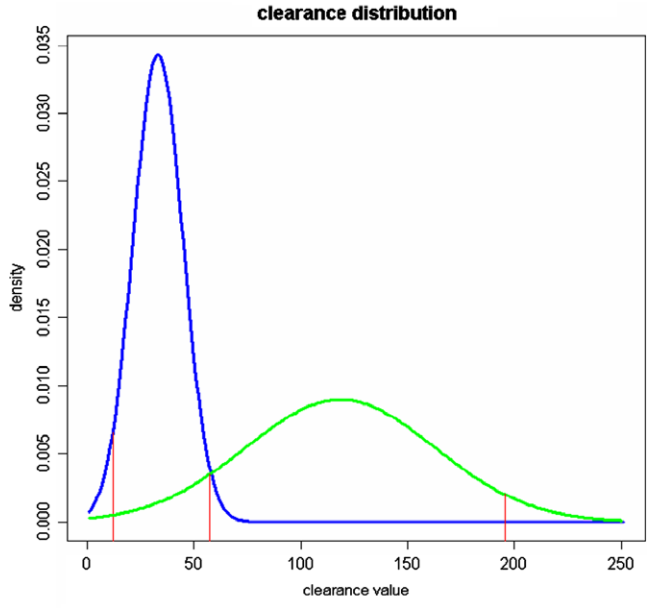

Fig. 3. The estimated clearance distribution. *Note:* The blue curve shows systemic clearance; the green curve shows oral clearance. The 95% confidence interval is marked on each curve using vertical lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Based on the distributions, the unknown type of clearance data were classified into oral clearance or systemic clearance, and outliers were removed. After the evaluation process, the final mined MDZ clearance data was shown in the second row of Table 4. The evaluation removes most of the false positive data. The left false positive data are comparable to the true clearance data, and they cannot be identified as outliers. Some true MDZ clearance data, labeled bold in the first row of Table 4, are considered as outliers by the evaluation. The clearance data before and after the evaluation are shown in Fig. 4. Obviously the meta-analysis can efficiently classify the data and remove the outliers.

3.1.5. Performance comparisons with DiDB database

To better evaluate our literature mining method, we compare the extracted MDZ clearance data with those from DiDB database. DiDB [26] is the most complete PK database so far, which is built manually. DiDB MDZ clearance data are downloaded and are compared with the mining MDZ clearance data. Table 5 lists detailed comparisons. DiDB [26] provides 11 PK relevant articles for MDZ. We read through their abstracts and find only six clearance data from healthy subjects. While the mining returns 170 PK relevant articles for MDZ, in which more than 70 clearance data are extracted from the abstracts. Therefore, the literature mining method possesses 70/6 = 11.6 times fold increase in information content, in addition to the benefits of the automatic data extraction.

The population mean and its standard error $(\theta \pm se)$ are calculated for DiDB clearance data and the mined clearance data. The true known population mean and standard error, which are calculated based on the manually extracted clearance data from relevant article abstracts, are given as a comparison benchmark. For the oral clearance, the benchmarker estimate is 83.6 ± 8.6 (L/h), while the DiDB and mining estimates are 58.3 ± 16.8 and 78.1 ± 6.0, respectively. Comparing to the benchmarker, the DiDB estimate is much more biased than our mining approach, 30.3% vs. 6.6%; and DiDB estimate's se is 2.8 times higher than our mining approach. For the systemic clearance, comparing to the benchmark, DiDB estimate's bias = (32.3–25.8)/32.3 × 100% = 20.1%, and mining estimate has a bias of 13.9%. DiDB estimate's se is 3.1 times higher than the se of our mining estimate.

One observation on the DiDB oral clearance data is the influence of the publication errors on the data analysis. PubMed PID 15470333 reported oral clearance for midazolam as 533 ± 759 ml/min by typo in the abstract. The correct value should be 1533 ± 759 ml/min in the full text. In the meta-analysis of our text mining, the influence of such error is eliminated by the outlier detection. However, DiDB database suffers from this type of publication error, and we suspect that DiDB only reads the abstract sometimes.

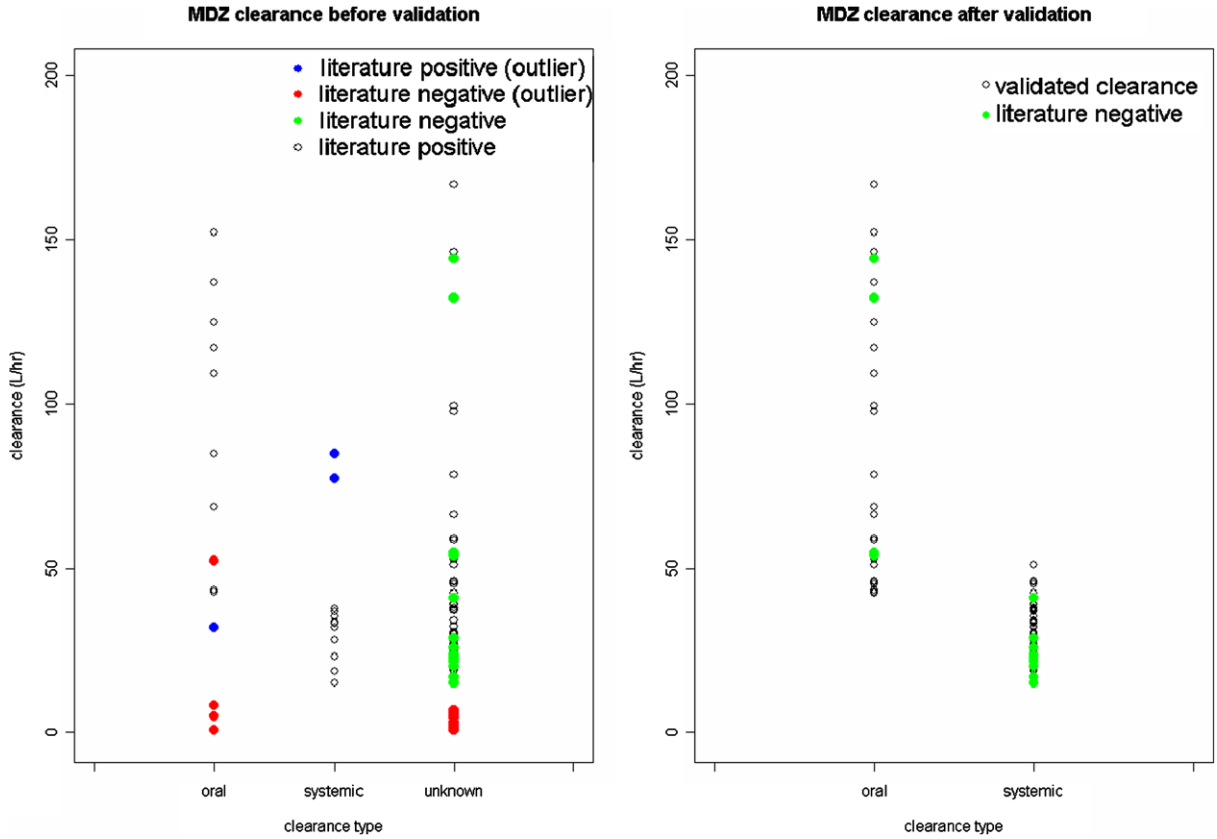

Fig. 4. MDZ clearance data. (a) Contains all mined MDZ clearance data before evaluation and outlier removal, and (b) contains the MDZ clearance data after evaluation outlier removal. The blue dots are true clearance data from MDZ PK relevant abstracts; the red and green dots are false MDZ clearance data, in which the red ones were removed by EM validation as outliers and green ones were not. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
MDZ clearance estimate comparisons among true known data, DiDB, and mining results.

	True known			DiDB			Mining		
	Abstract PK #	Relevant article #	$\theta \pm se$	Abstract PK #	Relevant article #	$\theta \pm se$	Abstract PK #	Relevant article #	$\theta \pm se$
Oral clearance	25	170	83.6 ± 8.6	2	11	58.3 ± 16.8 (88.4 ± 7.3)[a]	28	170	78.1 ± 6.0
Systemic clearance	50		32.3 ± 1.8	4		25.8 ± 3.1	59		27.8 ± 1.0

Note: This table shows how many PK relevant articles ("relevant article #") were are available, and how many clearance data ("abstract PK #") were extracted from their abstracts.

[a] The estimate of oral clearance after an outlier (publication bias in DiDB) is removed from the data set. This outlier is automatically removed in our mining approach.

3.2. Recall and precision

3.2.1. Validation data generation

The classical way to evaluate the performance of information retrieval is to check its recall and precision. In this case study, the quality of the entity template determines how well the MDZ PK relevant abstracts can be retrieved. However since the sample data (over 7000 abstracts) from PubMed search are too big to be handled manually for the recall and precision analyses, a subset of the abstracts are generated to estimate the performance of each literature mining step.

To build such a subset, one more keyterm "pharmacokinetics" is included into the PubMed search. This decreases the size of the result abstracts to 819, a reasonable number for the manual performance check. The results are shown in Fig. 5. The manual inspection of the 819 abstracts returns 164 PK relevant articles for drug MDZ.

3.2.2. Entity recognition

After applying the entity template, 220 out of the 819 abstracts are left in which 150 abstracts are truly relevant. The recall of this information retrieval step is 91% and the precision is 68% (Table 6).

3.2.3. Comparison with automatic abstract classification

To evaluate the power of this entity template, we compare the performance of template based abstract classification with an automatic classifier implemented using a support vector machine (SVM). Training data were established by dividing the 164 relevant abstracts into three groups with about 55 abstracts in each, then adding to each group 55 randomly selected irrelevant abstracts. The group which generates higher *F*-score was recorded as SVM (50). We applied a two-step process to determine proper features for SVM. First, a chi-square based feature

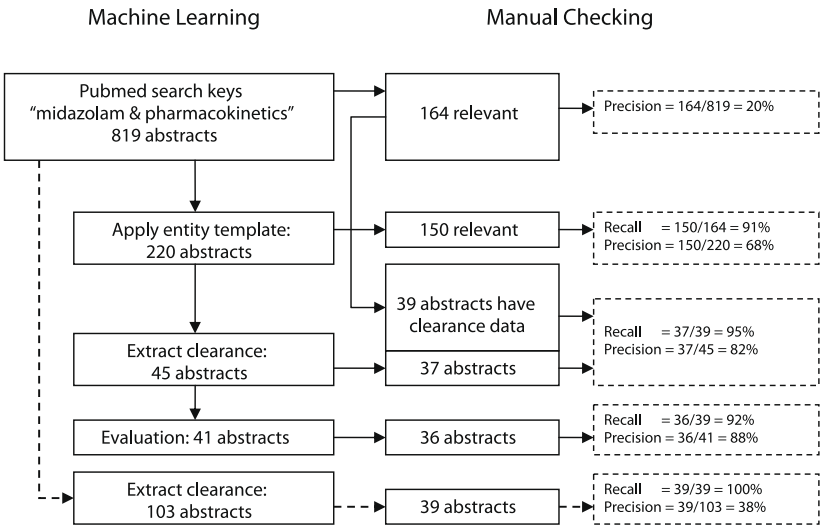

Fig. 5. Recall and precision performance analysis of the machine learning algorithm in a MDZ abstracts subset.

Table 6
Comparison between template and SVM methods on MDZ abstract relevance.

Method	Total	Query	TP	FP	FN	TN	Precision (%)	Recall	*F*-score	Accuracy
MDZ-relevance										
PubMed Query	NA	819	164	655	NA	NA	20.0	NA	NA	NA
PubMed query + entity template	819	220	150	70	14	585	68.2	91.5%	78.1%	89.7%
SVM (50)	819	159	110	49	54	496	69.2	67.1%	68.1%	74.0%
SVM (100)	819	277	142	135	22	520	51.3	86.6%	64.4%	80.8%

Note: The training data of SVM (50) contains 50 randomly selected relevant abstracts and 50 irrelevant; the training data of SVM (100) contains 100 randomly selected relevant and 100 irrelevant. (TP, FP, FN, and TN) stand for true positive, false positive, true negative, and false negative, respectively.

selection filter was used to retain all features with the *P*-value below threshold 0.05. Then, the remaining features went through a principle component analysis [27] for dimensionality reduction, which was set to keep a cumulative proportion 95% of the original features. The final features were fed into SVM for model training and classification. We also tried a second training data set, which was made up of 100 randomly selected abstracts from the 164 relevant articles and 100 randomly selected irrelevant abstracts. The SVM^{light} [18] was implemented with different kernels, and the best performance was shown in Table 6. The precision/recall is measured on information retrieval, finding relevant articles out of the test set of abstracts. SVM (50) has slightly higher precision than our entity template in identifying MDZ relevant abstracts (69.2% vs. 68.2%), but much worse recall (67.1% vs. 91.5%). Hence SVM (50)'s *F*-score is lower than entity template (68.1% vs. 78.1%). On the other hand, SVM (100) generates reduced precision, 51.3%, and improved recall, 86.6%. Its *F*-score becomes even worse, 64.4%. Overall, our literature mining approach out-performs SVM. As the features used for SVM (100) covers only 56% of the template features, to further explore the difference between the template based classification with SVM, we combined the features from the chi-square feature selection with template features and fed them directly into SVM for model training and classification. The inclusion of extra features did not improve the performance of SVM at all. It is possible that SVM has not been able to model many non-linear or interactive relationships, which were introduced into the templates implicitly.

3.2.4. Information extraction

For the clearance data, the manual inspection proves 39 out of the 164 relevant abstracts containing MDZ clearance numerical values (clearance relevant). Our information extraction step recognizes 45 abstracts as clearance relevant, in which 37 are true. Hence, the recall rate for clearance data extraction is 95% and the precision is 82%. The same information extraction rules are also applied directly to the starting 819 abstracts (Table 7). Without the application of entity template, the precision drops from 82% to 38%, and *F*-score reduces from 88% to 55%.

Table 7
clearance extraction with and without entity template.

Method	Total	Query	TP	FP	FN	TN	Precision (%)	Recall (%)	*F*-score (%)	Accuracy (%)
Clearance-relevance										
PubMed query + CL extraction	819	103	39	64	0	716	37.9	100.0	54.9	92.2
PubMed query + entity template + CL extraction	819	45	37	8	2	772	82.2	94.9	88.1	98.8
PubMed query + CL extraction + outlier evaluation	819	73	39	34	0	746	53.4	100.0	69.6	95.8
PubMed query + entity template + CL extraction + outlier evaluation	819	41	36	5	3	775	87.8	92.3	90.0	99.0

Table 8
CL data extraction on more drugs: DiDB vs. literature mining.

Drug name	DiDB			Mining			Comparisons		
	N	*n*	*p* (%)	*N*	*n*	*p* (%)	Coverage	*n*-FC	*p*-FC
Information content comparison									
Triazolam	37	6	16	11	11	100	100%	1.83	6.25
Alprazolam	44	8	18	22	18	82	100%	2.25	4.55
Nifedipine	41	5	12	22	11	50	100%	2.2	4.12
Nitrendipine	2	0	0	5	3	60	N/A	Inf	Inf
Diazepam	3	3	100	4	3	75	100%	0	−0.25
Amlodipine	4	1	25	5	4	80	100%	4.0	3.2
Nitrendipine	2	2	100	5	3	60	100%	1.5	−0.40

Note: N, total number of reported abstracts in DiDB; and number of extracted abstracts from text mining.
n, clearance relevant abstracts.
p, precision = *n*/*N*.
Coverage, the percentage of DiDB clearance relevant abstract covered by text mining approach.
n-FC, fold-change from DiDB to mining in clearance relevant abstracts, *n*.
p-FC, fold-change from DiDB to mining in precision, *p*.

3.2.5. Evaluation

The meta-analysis evaluation removed most outliers and false positive values. After this step, the clearance data from 41 abstracts are left and 36 of the abstracts are true MDZ clearance relevant. The recall rate becomes 92% and the precision is improved to 88%. Similarly, without the entity template step, both the *F*-score and precision drop significantly (Table 7), from (90%, 88%) to (70%, 53%).

3.2.6. Comparison of MDZ data mining and its validation analysis

Figs. 2 and 5 show the PK information comparison between single PubMed search keyterm (“midazolam”) and two keyterms (“midazolam” and “pharmacokinetics”). Though the PubMed search returns much more abstracts using a single keyterm than using two keyterms (7129 vs. 819), only six more relevant abstracts are found in the single term search results (170 vs. 164). The difference of the number of clearance relevant abstracts is also six (45 vs. 39).

3.3. Information content comparison with DiDB

Table 5 suggests that our literature mining approach collects 11 times more MDZ clearance data than the manually curated DiDB database contains. To test the portability of our literature mining method, we tried it on 7 other Cytochrome P450 3A Subfamily drugs and extracted their clearance data from PubMed abstracts as for Midazolam. The same drugs were also searched in DiDB database (September 2008), and clearance data was also analyzed. The comparison was shown in Table 8. Among 5 out of 7 drugs, comparing to DiDB, literature mining generates 1.83- to 4.0-fold more information contents in CL, and precision increases from 3.2 to infinite fold higher. Among those two drugs that DiDB outperforms literature mining, our approach only misses totally two abstracts.

4. Conclusions

In this paper, we tested the feasibility of literature mining on drug PK parameter numerical data by designing a sequential mining strategy. Firstly, an entity template library is built to retrieve pharmacokinetics relevant articles. Then a set of tagging and extraction rules are applied to retrieve PK data from the article abstracts. To estimate the PK parameter population mean and between-study variance, a linear mixed meta-analysis model and an E–M algorithm are established to describe the distribution of PK parameters. Finally, a leave-one-out cross-validation procedure is developed to ascertain false-positive mining results based on the linear mixed meta-analysis model. Using this approach to mine MDZ PK data, an *88% precision rate* and *92% recall rate* are achieved. A conventional data mining approach, SVM, is compared to our method. Its performance is nowhere near our approaches. Our text mining approach recollects *11 times* more MDZ clearance data than a manual accumulated DiDB database has. Interestingly, it also identifies a publication error of midazolam clearance data, which cannot be assessed in the DiDB database. In addition, we also establish the first validation set for more general data mining methodology development for PK data.

With extensive evaluation, it reveals that our literature mining’s outstanding recall and precision performance is largely due to the well constructed entity template. This entity template outperforms SVM in identifying MDZ relevant abstracts. We further investigate our literature mining approach in 7 more CYP3A substrate drugs. Among five out of seven drugs, comparing to DiDB, literature mining generates 1.83- to 4.0-fold more information contents in CL, and precision increases from 3.2 to infinite fold higher. Among those two drugs that DiDB out-performs literature mining, our approach only misses totally two abstracts. Therefore, from the information content point of view, our data mining approach out-performs DiDB. At the meantime, since we implemented statistical model based evaluation strategies for the mining data, our integrated approach can identify outliers for quality control (QC). As a side production of QC, we provide not only population PK parameter estimates of PK parameters, but also their variations estimates. These features are not available in DiDB. Most importantly, DiDB could have update lag, while our approach is in real time.

The performance of our proposed mining strategy on MDZ PK study is so promising that this feasibility study encourages us for more case studies. Although the entity recognition template is very general in our current setting, its performance on the other drugs needs more assessment. Please note that MDZ usually serves as a CYP3A probe drug in pharmacokinetics studies, its information is in general richer than the other non-probe drugs.

For the mining of pharmacokinetics data, one important issue is the drug name recognition. In this study, the co-existence of other drugs was tagged for syntax analysis in the information extraction step. Furthermore, their interactions with the object drug can be considered to provide extra information for the mining. For example, inducers (e.g. rifampin) of midazolam increase its clearance and inhibitors decrease it. This brings in a valuable guidance for clearance data extraction if the inducer/inhibitor of the object drug can be tagged correctly. In this paper, we built a drug name dictionary based on FDA databases. We are also evaluating other resources for more drug information, e.g. DrugBank (http://www.drugbank.ca), PubChem (http://pubchem.ncbi.nlm.nih.gov), Drugs.com and RxList (http://www.rxlist.com).

The research presented in this paper also inspires our next step work on the application of the literature mining technique, which is full text based mining on drug PK data. As full texts usually contain much more PK numerical data than abstracts, we should be able to get more useful information. For example, the definition of “population PK parameter” in the data evaluation step is only under the statistical consideration. Whether the population represents the “world population” is highly debatable. In principle, it should be determined by the racial, gender, and study location compositions from different studies. This annotation information is not always available in the abstracts, but is available in the full text. Thus the full text articles would be the actual resource for building the drug PK database. The current work in this paper serves a very much needed starting point.

References

[1] Woosley RL, Cossman J. Drug development and the FDA's critical path initiative. Clin Pharmacol Ther 2007;81:129–33.
[2] Veit M. New strategies for drug development. Berl Munch Tierarztl Wochenschr 2004;117:276–87.
[3] D'Andrea G, D'Ambrosio RL, Di Perna P, Chetta M, Santacroce R, Brancaccio V, et al. A polymorphism in the VKORC1 gene is associated with an interindividual variability in the dose-anticoagulant effect of warfarin. Blood 2005;105:645–9.
[4] Kirchheiner J, Brockmoller J. Clinical consequences of cytochrome P450 2C9 polymorphisms. Clin Pharmacol Ther 2005;77:1–16.
[5] Badagnani I, Castro RA, Taylor TR, Brett CM, Huang CC, Stryke D, et al. Interaction of methotrexate with organic-anion transporting polypeptide 1A2 and its genetic variants. J Pharmacol Exp Ther 2006;318:521–9.
[6] Hung SI, Chung WH, Jee SH, Chen WC, Chang YT, Lee WR, et al. Genetic susceptibility to carbamazepine-induced cutaneous adverse drug reactions. Pharmacogenet Genom 2006;16:297–306.
[7] Food and Drug Administration. Innovation or stagnation: challenges and opportunity on the critical path to new medical products; 2004. Available from: http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.
[8] Food and Drug Administration. Critical path opportunity list; 16 March 2004. Available from: http://www.fda.gov/oc/initiatives/criticalpath/reports/opp_list.pdf.
[9] Food and Drug Administration. Critical path opportunity report; 2006. Available from: http://www.fda.gov/oc/initiatives/criticalpath/projectsummary/consortium.html.
[10] Lalonde RL, Kowalski KG, Hutmacher MM, Ewy W, Nichols DJ, Milligan PA, et al. Model-based drug development. Clin Pharmacol Ther 2007;82(1):21–32.
[11] Chang M, Kenley S, Bull J, Chiu YY, Wang W, Wakeford C, et al. Innovative approaches in drug development. J Biopharm Stat 2007;17(5):775–89.
[12] O'Neill RT. FDA's critical path initiative: a perspective on contributions of biostatistics. Biom J 2006;48(4):559–64.
[13] Chien JY, Friedrich S, Heathman MA, de Alwis DP, Sinha V. Pharmacokinetics/pharmacodynamics and the stages of drug development: role of modeling and simulation. AAPS J 2005;7(3):E544–59.
[14] Pustejovsky J, Castaño J, Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. In: Proceedings of the seventh pacific symposium on biocomputing world scientific, Hawaii; 2002. p. 362–73.
[15] Thompson DF, Williams NC. Tracking the growth of drug therapy literature using PubMed. Drug Inform J 2007;41(4):449–55.
[16] Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Comput Biol 2003;10(6):821–55.
[17] Jensen LJ, Saric J, Brok P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006;7:119–29.
[18] Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A, editors. Advances in kernel methods – support vector learning. MIT-Press; 1999.
[19] Yeh AS, Hirschman L, Morgan AA. Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. ISMB (Suppl Bioinform) 2003:331–9.
[20] Hersh W, Cohen AM, Roberts P, Rekapalli HK. TREC 2006 genomics track overview. In: TREC 2006 notebook; 2006. p. 68.
[21] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinform 2005;6(Suppl. 1):S1.
[22] Pentikäinen PJ, Valisalmi L, Himberg JJ, Crevoisier C. Pharmacokinetics of midazolam following intravenous and oral administration in patients with chronic liver disease and in healthy subjects. J Clin Pharmacol 1989;29(3):272–7.
[23] opennlp. Available from: http://opennlp.sourceforge.net/index.html.
[24] Reynar JC, Ratnaparkhi A. A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the fifth conference on applied natural language processing; 1997.
[25] Porter MF. An algorithm for suffix stripping. Program 1980;14(3):130–7.
[26] Metabolism and transport drug interaction database. Available from: http://depts.washington.edu/didbase; 1999–2003 [accessed: 03.07]. Copyright University of Washington.
[27] Wall ME, Rechtsteiner A, Rocha LM . Singular value decomposition and principal component analysis. In: Berrar DP, Dubitzky W, Granzow M, editors. A practical approach to microarray data analysis. Norwell, MA: Kluwer; 2003. p. 91–109.
[28] Yu M, Kim S, Wang Z, Hall S, Li L. A Bayesian meta-analysis on published sample mean and variance pharmacokinetic data with application to drug–drug interaction prediction. J Biopharm Stat 2008;18(6):1063–83.